Cell, Volume *129*

Supplemental Data

Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs

John L. Rinn, Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha A. Brugmann, L. Henry Goodnough, Jill A. Helms, Peggy J. Farnham, Eran Segal, and Howard Y. Chang

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Tiling Array Design

Sequences for each genomic region (**Tables S1 and S2**) was retrieved from UCSC human genome sequence version 17; repetitive sequence elements were removed by RepeatMasker. Four hundred thousand 50mer probes were printed using Nimblegen technology (Nimblegen Inc., Madison, WI); each probe overlapped the previous probe by 45 nucleotides, thereby reporting on five unique bases. We used the algorithm ArrayOligoSelector (Bozdech et al., 2003) to computationally test each probe for (i) self annealing; (ii) thermodynamics of target hybridization; and (iii) selectivity against the remainder of the genome. We found that because the paralogous HOX sequences had nucleotide variations sprinkled throughout, the majority (>98.5%) of our probes had significant free energy differences (>30 kcal/mol) from the next best match in the genome such that these probes can be considered mono-specific. We confirmed the specificity of the tiling array for distinguishing highly related sequences within the HOX loci by hybridization with single cDNA products. For instance, the HOXA13 cDNA hybridized to only the HOXA13 regions of the tiling array exactly matching the two exons, but not to the similar paralogs HOXB13, HOXC13, or HOXD13 (**Figure S2**).

RNA isolation and probe preparation

Eleven primary human fibroblast cultures were established from skin and internal organ donors as described (Rinn et al., 2006) from discarded normal tissue during surgical repair or from autopsy using a skin punch. The anatomic sites were mapped using a preset diagram based on bony landmarks of the donors. Fibroblasts were isolated from keratinocytes and endothelial cells, and their lineage confirmed by immunofluorescence for vimentin positivity and negative staining for cytokeratin, desmin, CD31, GFAP, and CD11b. Total RNA was purified from each cell culture using Trizol according to the manufacturer's instructions (Invitrogen). Tiling array probes were prepared from total RNA, which was amplified (aRNA) using Message Amp II (Ambion), The aRNA was reverse transcribed with random decamers using RETROscript Kit (Ambion), and the resulting cDNA was labeled and hybridized to Nimblegen arrays as described (Squazzo et al., 2006).

Identification of discrete transcribed regions in the HOX locus. To identify discrete transcribed regions from the raw hybridization intensity measurements in an unbiased way, we first computed the weighted average log-intensity at each 100 bp consecutive genomic window in the HOX locus, where this computation was done separately for each of the 11 microarrays. Specifically, the contribution of a probe to a window with which it has an overlap of at least five bp, is the log-intensity of the probe multiplied by the size of the overlap between the probe and the window. The weighted average of a probe is then the sum of all contributions of its overlapping probes, divided by the total size of the overlap between the window and all of its contributing probes. Next, for each microarray experiment, we automatically identified an intensity threshold between expressed and non-expressed probes, by identifying the threshold for which the t-test between its two resulting groups of expressed non-expressed probes gives a minimal p-value. For each microarray, we then defined its expressed 100 bp windows from above as those windows whose weighted average intensity was above the expressed threshold for the microarray. Finally, we combined the expressed windows across all microarrays into one collection, and defined every set of overlapping expressed windows as a discrete transcribed region, whose boundaries are defined by the lowest and highest genomic coordinate of all of its constituent expressed windows. Transcribed regions that contained locations of known transcription start sites (TSS) of HOX genes were separated into two transcribed regions, with one transcribed region fully contained within the known HOX transcript and the other fully contained within a HOX intergenic region.

ncRNA expression ratios, correlation with neighboring HOX genes, and correlation with PRC2 chromatin domains . We defined the expression ratio of each of the above discrete transcribed regions in every microarray sample, as the weighted average of the log-intensity of all of its overlapping probes, minus the average log-intensity of all of the probes in the microarray. For each ncRNA a Pearson correlation was calculated by comparing the expression ratios of each ncRNA in all 11 samples to the expression ratios of the 5' HOX gene in all 11 samples and the 3' HOX gene in all 11 samples. To determine correlations between ncRNAs and flanking HOX genes that are significantly (P < .05) higher to the 3' or 5' HOX gene we performed a power calculation. For 11 samples we acquire 60% power when the difference (delta) between the 5' and 3' HOX gene Pearson correlation value to a given ncRNA was greater than 0.6 (**http://calculators.stat.ucla.edu/ powercal**). Chi-square tests for UBX and ANTP evolutionary origins (Figure 3) and chromatin mark and RNA comparisons (Figure 4) were performed (http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html).

Hierarchical clustering of ncRNA expression values was performed by CLUSTER (Eisen et al., 1998). Correlation to ncRNAs to the 3' HOX gene was performed by Pearson correlation of the expression values of each ncRNA in all 11 samples to the expression values of each 3' HOX gene in all 11 samples. Relative distance of each ncRNA to the 3' HOX gene was calculated by the absolute difference of the 3' coordinate of the ncRNa boundary from the 5' coordinate of the HOX gene UCSC known gene annotation and divided by the total distance between the 3' coordinate of the 5' HOX gene from 5' coordinate of the 3' HOX gene.

EST analysis was performed using the UCSC genome browser (http://genome.ucsc.edu/). Coordinates for each ncRNA was submitted to the browser and was interrogated for a previously detected EST or mRNA. If detected information was recorded for the strand for each EST relative to the surrounding HOX genes as well as the length of extension for each ncRNA detected by DNA tilling array. A total of 25 ncRNAs located in the region of HOXC4 and HOXC6 were not included in this analysis as current annotation show a large overlapping intron of HOXC6 that extends into HOXC4 and encompasses these 25 ncRNAs; annotations in this region were unclear and thus excluded.

To correlate HOX and ncRNA expression with Suz12 occupancy, we classified each transcript as induced (or repressed) by determining its expression level relative to the mean across the 11 samples. Transcripts above the mean were deemed induced in those samples. Chromatin domains of Suz12 occupancy from ChIP-chip data were visually scored, and the association between transcription and Suz12 occupancy was determined using a 2X2 chi-square test.

Reverse Transcription PCR

The RT-PCR in Figure 2 and Figure S3 was performed using the RetroScript kit (Ambion) and PCR was performed using the following primers : tar-HOXD10-8 (Forward CCTTAATTTCCCTGCAAACG, Reverse TGAAGGTGTAAGGCTGCACA), tar-HOXD3-39 (ForwardCCCACGCATCTCTATTTGGTCR CATACCAAATGCCACACAGG, Reverse CCTCCAAGCTGAGAAGGAGA), tar-HOXD1-46 (Forward

CCCACGCATCTCTATTTGGT, Reverse CAGCACAAAGGAACAAGGAA), tar-HOXB4-168 (Forward AAGCTGTGACAGAGTAAGGGAAA, Reverse GGGTTGTTTTTGTGTTGTTCC), tar-HOXA11-93 (Forward CCTTAATTTCCCTGCAAACG, Reverse TGAAGGTGTAAGGCTGCACA), tar-HOXA6-72 (Forward CATTGCGGAGGGCATTGG, Reverse GTACGCCCTGATGTTTCC and HOTAIR (Forward GAGAACGCTGGAAAAACCTG, Reverse TTGGGGAAGCATTTTCTGAC). The RT-PCR data was considered in agreement with the array expression values if a band was present and the expression value for the ncRNA in that sample was greater than the mean (red) or in the absence of a band the mean expression was below average for that ncRNA (green). Strand specific RT-PCR of HOTAIR was performed using primers F1 (GGGGCTTCCTTGCTCTTCTTATC), R1 (CTGACACTGAACGGACTCTGTTTG), F2 (GGTAGAAAAAGCAACCACGAAGC), R2 (ACATAAACCTCTGTCTGTGAGTGCC), and F3 (CGGAGGTGCTCTCAATCAGAAAG) for strand specific cDNA synthesis.

The quantitative PCR performed in figure 5 was performed by Brilliant® SYBR® Green® QRT-PCR (Stratagene) according to manufacturer’s instructions using 50ng of total RNA. HOTAIR primers F2 and R2 were used with a 55 C annealing temperature for 40 cycles. All HOTAIR levels were normalized to GAPDH (Forward CCGGGAAACTGTGGCGTGATGG, Reverse AGGTGGAGGAGTGGGTGTCGCTGTT).

Identification of common DNA sequence motifs in transcriptionally active regions (ncRNAs). To identify common DNA sequence motifs in ncRNAs, we used a discriminative motif finder that we previously developed (Segal et al., 2003). Specifically, for each gene set considered, the discriminative motif finder searches for motif sequences that are commonly present in the ncRNA set of interest but are not present in all other transcribed regions of the HOX loci.

RNA secondary structure analysis.

Secondary structure analysis was performed using Mfold web server version 3.2 by Zuker and Turner (http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi). Input comprised of 200 nucleotide sequence fragments at every 50th nucleotide spanning the entire HOTAIR sequence.

SUPPLEMENTAL FIGURES

Rinn et al. Supp. Fig. 1

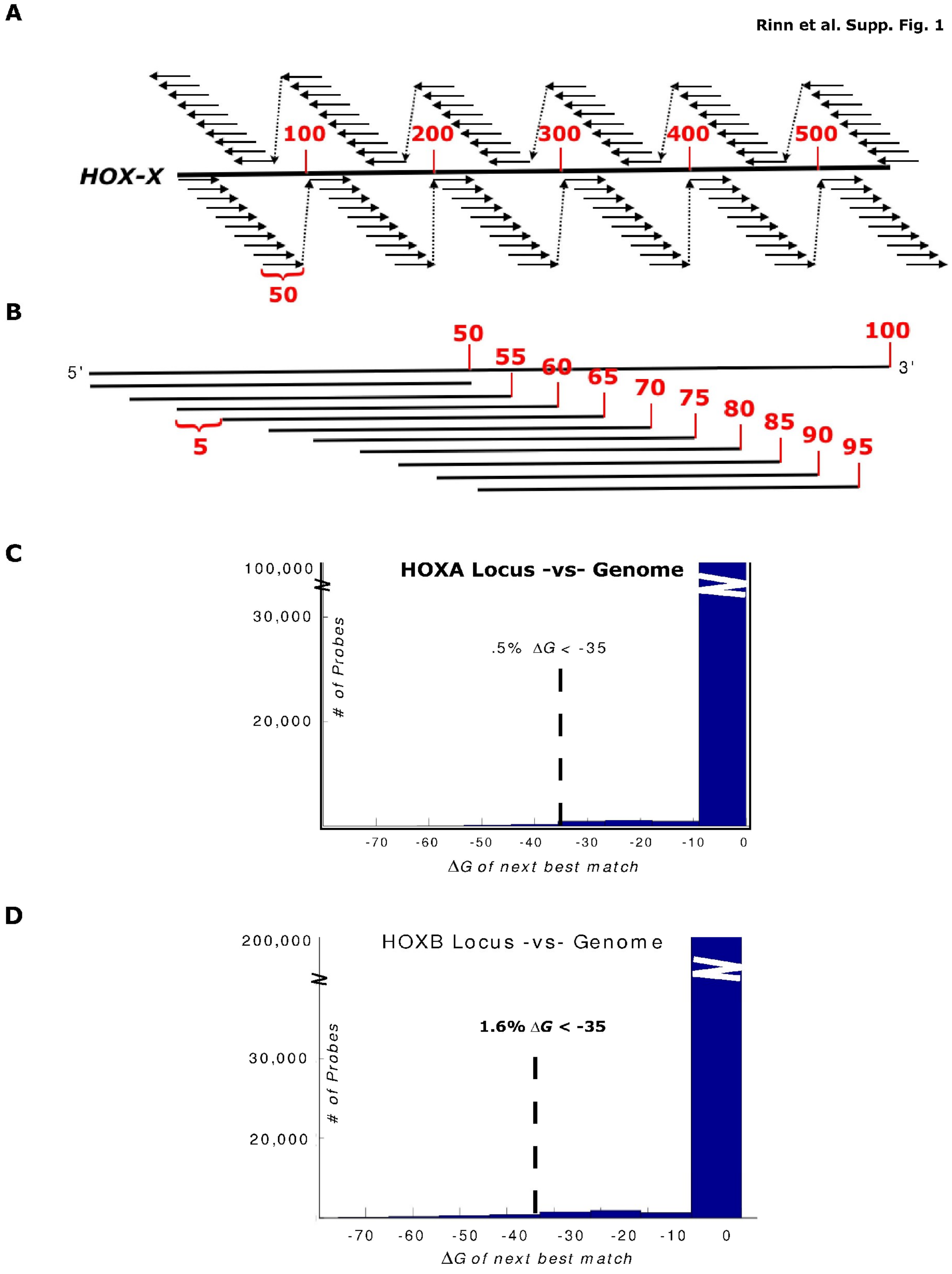

Figure S1. HOX tiling microarray. Design strategy and computational analysis of HOX tiling array probes.

(A) Sequences for each genomic region (Table S1) were retrieved from UCSC human genome sequence version 17; repetitive sequence elements were removed by RepeatMasker. The remaining non-repetitive sequence was tiled using 50mer oligonucleotides.

(B) Probes were selected to overlap 45 of 50 bp with the previous probe to cover each unique 5bp of non-repetitive sequence.

(C) Array Oligo Selector 3.8 (AOS3.8) analysis of cross-hybridization potential of each probe to other regions in the human genome. AOS3.8 blasts each probe against the entire human genome sequence to find all other sites in the genome with significant similarity and calculates a difference in the thermodynamic binding energy of the probe to its intended target versus the next best target in the human genome. Probes with binding energies lower than –35 have been shown to have potential to cross hybridize to other sequences in the human genome (Bozdech et al., 2003). Plotted on the X-axis are probes bined by the difference in binding energy and the Y-axis indicated the number of probes in each bin; only a small number of scattered probes did not fulfill the thermodynamic criteria. For instance, 0.5% and 1.6% of probes representing the HOXA (C) and HOXB (D) clusters respectively had potential to cross-hybridize.

Rinn et al. Supp. Fig. 2

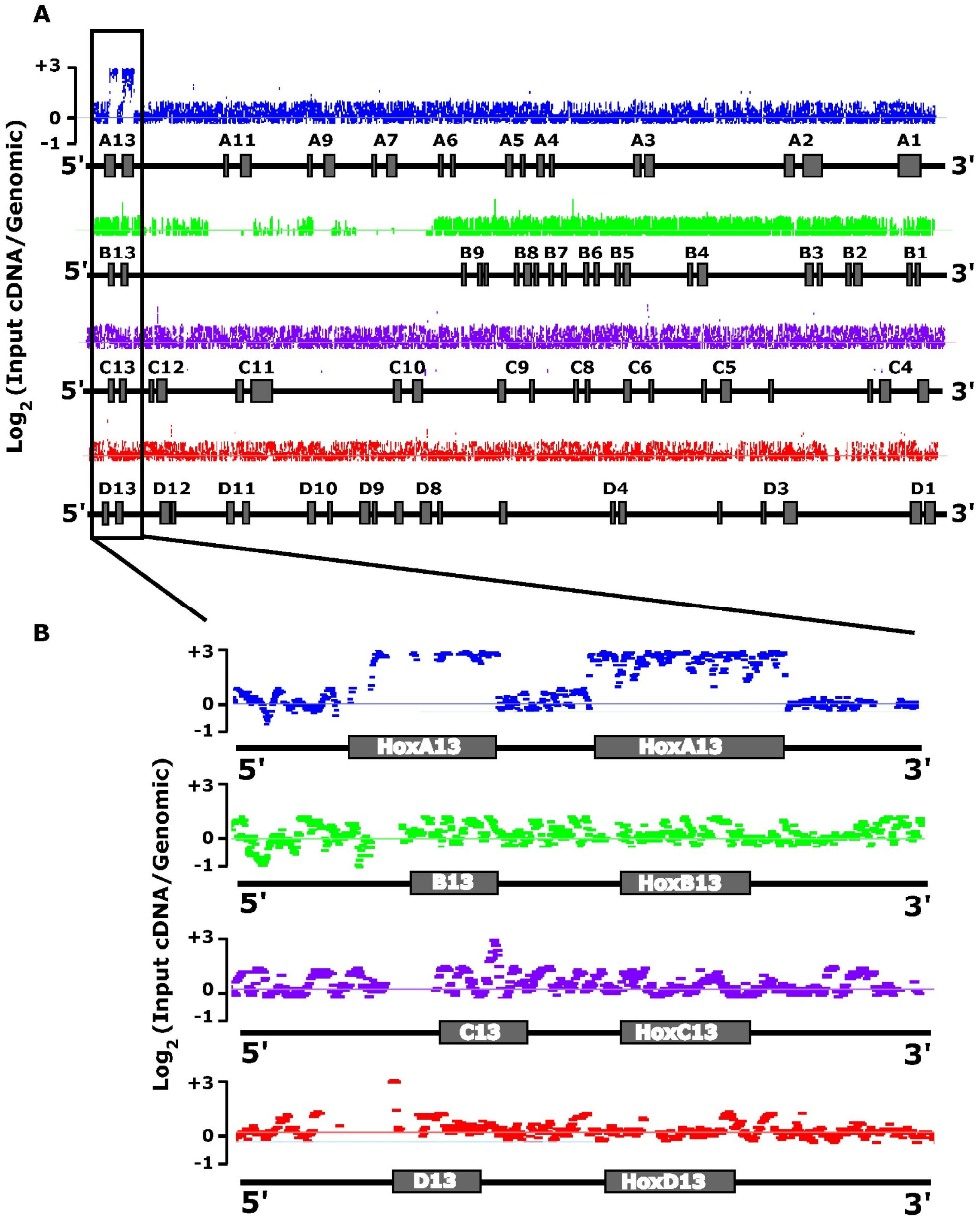

Figure S2. Experimental validation of HOX array probe specificity. A full-length clone of HOXA13 cDNA was labeled with Cy3 and hybridized to the array. Genomic DNA was labeled with Cy5 and hybridized to the array as a reference.

(A) The probe intensities of probes representing all four human HOX loci (HOXA-blue, HOXB-green, HOXC-purple, HOXD-red) are displayed as log base 2 ratio of the (HOXA13 cDNA/Genomic DNA) from +3 (8-fold) to –1 (.5 fold).

(B) Probe intensities of the four genes in paralogous group 13. Any cross-hybridization, such as from the highly conserved sequencing encoding the homeodomain, would have been detected as contiguous signal intensities on other HOX genes. No such regions were experimentally observed.

Rinn et al. Supp. Fig. 3

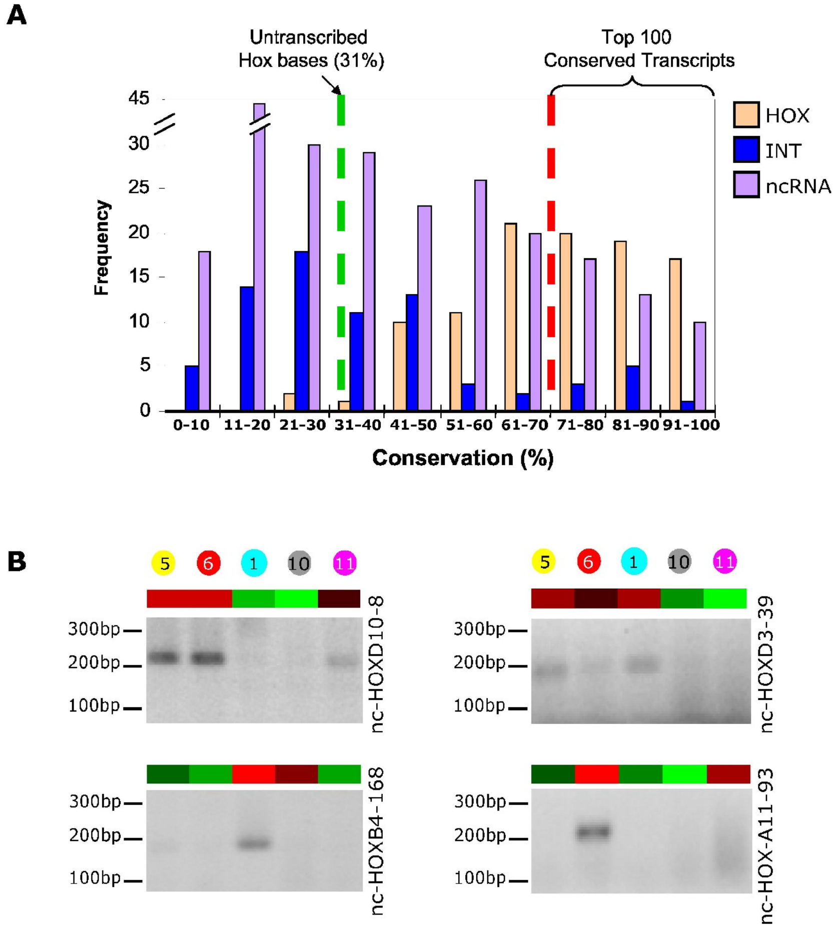

Figure S3. Evolutionary conservation and experimental validation of ncRNAs.

(A) Shown is the histogram of percentage conservation of HOX, INT, and ncRNA sequences across seven vertebrate species. The conservation of each base of the human HOX loci in the genomes of mouse, rat, dog, opossum, chicken, frog (*Xenopus tropicalis*), puffer fish (*Tetraodon nigroviridis*) was downloaded from the UCSC Genome Browser. For each transcript, we averaged the conservation value of all of its bases to produce a single conservation value. Notably, the average conservation per base of the following regions are: 0.69 for HOX genes, 0.44 for ncRNAs, 0.31 for INTs, and 0.31 for non-transcribed sequences. The averaged conservation value of the entire HOX loci is 0.39. Thus, the unbiased analysis of transcription in the HOX loci identified regions that are more conserved than would be expected by chance alone.

(B) RT-PCR validation of twenty predictions of ncRNA expression. Top: Tiling microarray data; bottom: RT-PCR. Fold difference in ncRNA expression predicted by tiling array in indicated by color intensity as shown on the scale bar.

Rinn et al. Supp. Fig. 4

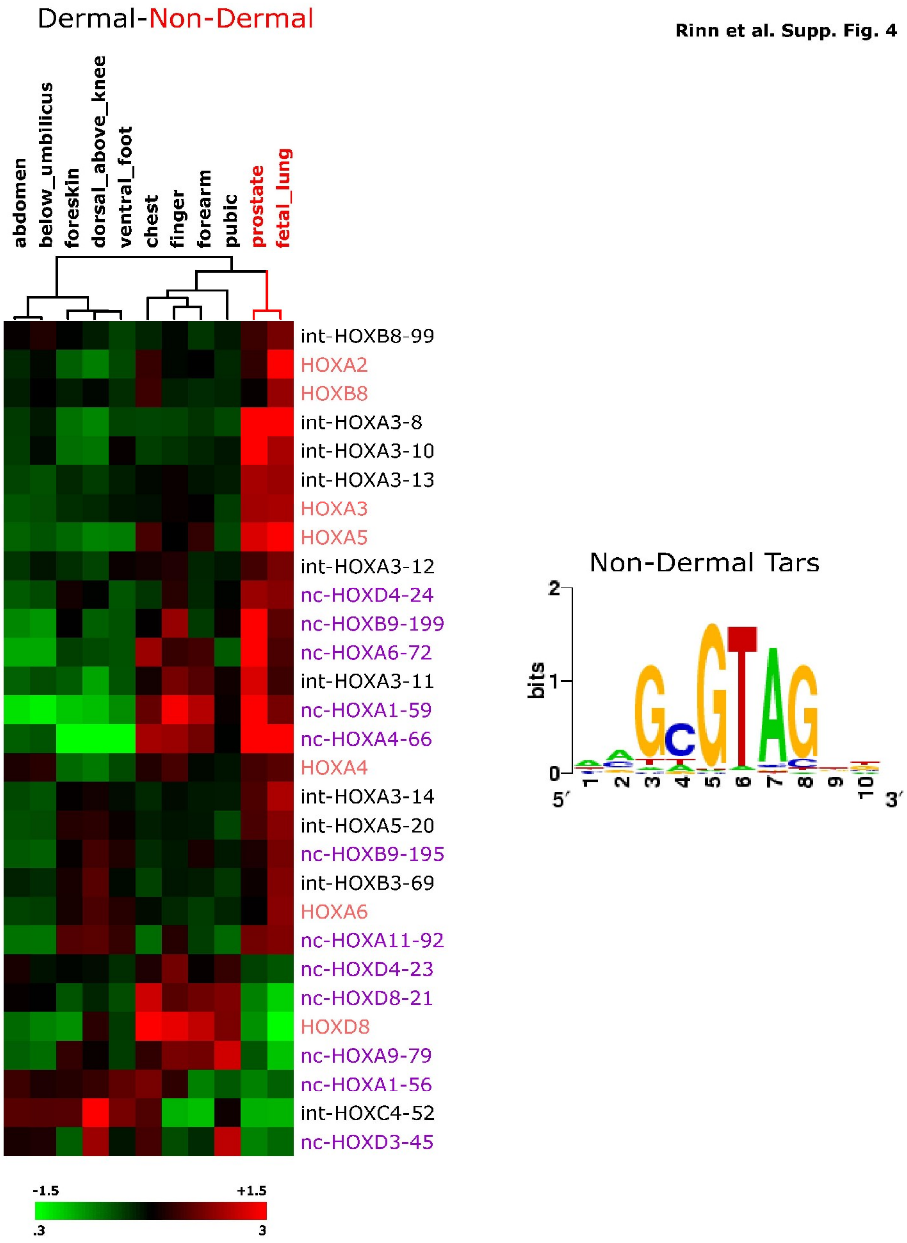

Figure S4. HOX transcripts exhibiting differntial expression between dermal and nondermal samples. A total of 7 HOX genes and 12 ncRNAs were differentially expressed between dermal and nondermal primary fibroblasts. The expression level of each transcript is represented relative to its median value in all 11 samples; expression levels above and below the global median are denoted by shades of red or green, respectively. The color scale encompasses a

range from 3 fold to .3 fold relative to global median transcript level for each gene (+1.6 to –1.6 logs on log base 2 scale). The samples were hierarchically clustered by transcript expression values and transcripts were hierarchically clustered by similarity across samples.

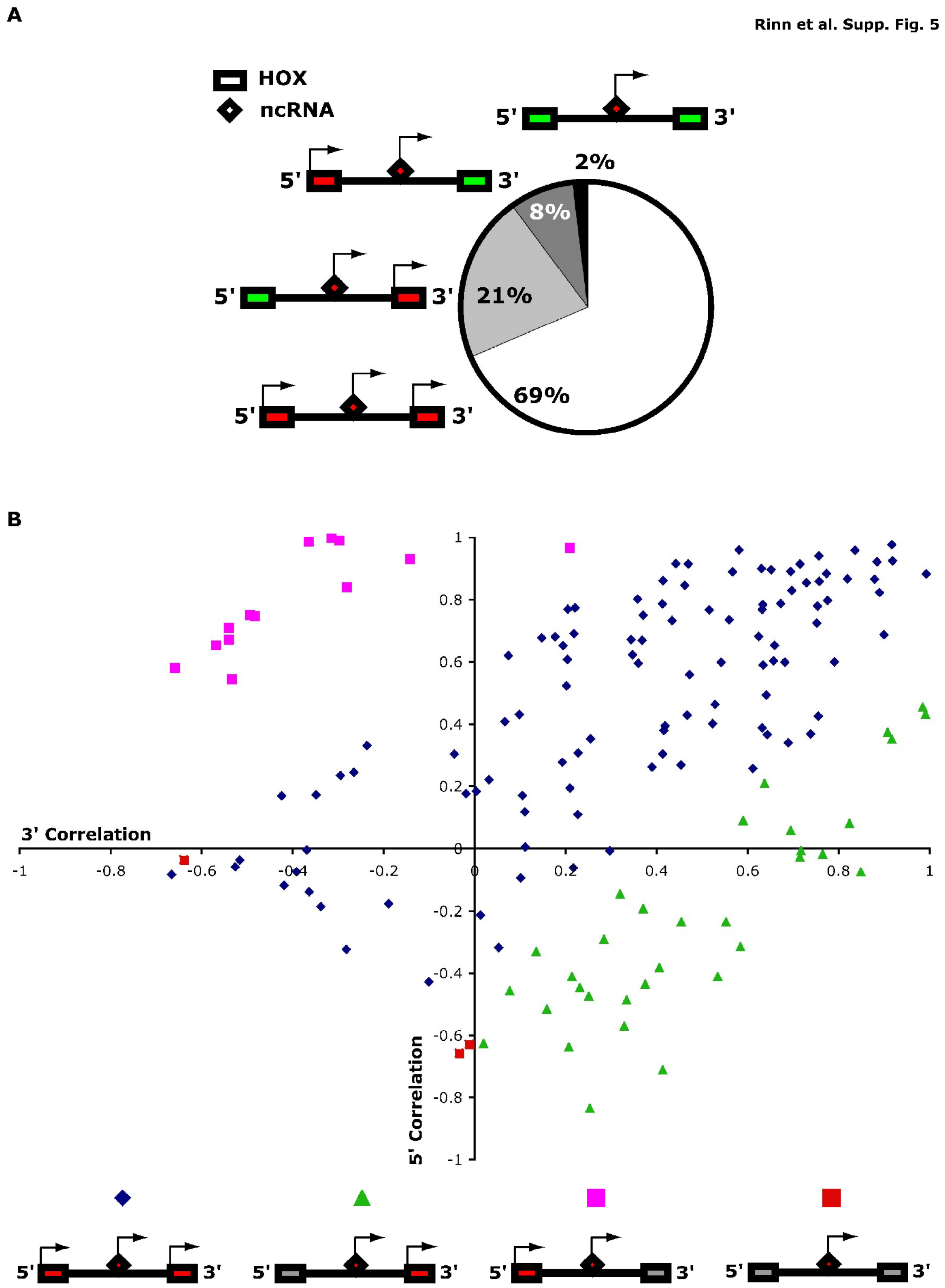

Figure S5. Correlation of ncRNA to 5’ and 3’ HOX gene expression patterns.

(A) Four patterns of ncRNA regulation and the percentage of ncRNAs in each category: i) ncRNA expression pattern is highly correlated (Pearson correlation) to both the 5’ and 3’ HOX gene expression pattern; ii) ncRNA expression pattern is significantly more correlated to the 3’ HOX gene; iii) ncRNA expression pattern is significantly more correlated to the 5’ HOX gene; iv) ncRNA is not significantly correlated to the 5’ nor 3’ HOX gene expression patterns.

(B) The Pearson correlation of each ncRNA expression profile with the expression profile of the flanking 5’ and 3’ HOX gene. The Pearson correlation value of each ncRNA expression profile in all 11 samples with the flanking 3’ HOX gene is plotted on the X-axis and the Pearson correlation value of each ncRNA and the 5’ HOX gene is plotted on the Y-axis. Blue diamonds: ncRNAs with positive correlation with both the 3’ and 5’ HOX genes; pink squares: ncRNAs with significantly higher (P < .05) correlation to the 5’ HOX gene; green triangles: ncRNAs with significantly higher (P < .05) correlation to the 3’ HOX gene; red squares: ncRNAs that are not correlated in expression to both the 5’ and 3’ HOX genes.

Rinn et al. Supp. Fig. 6

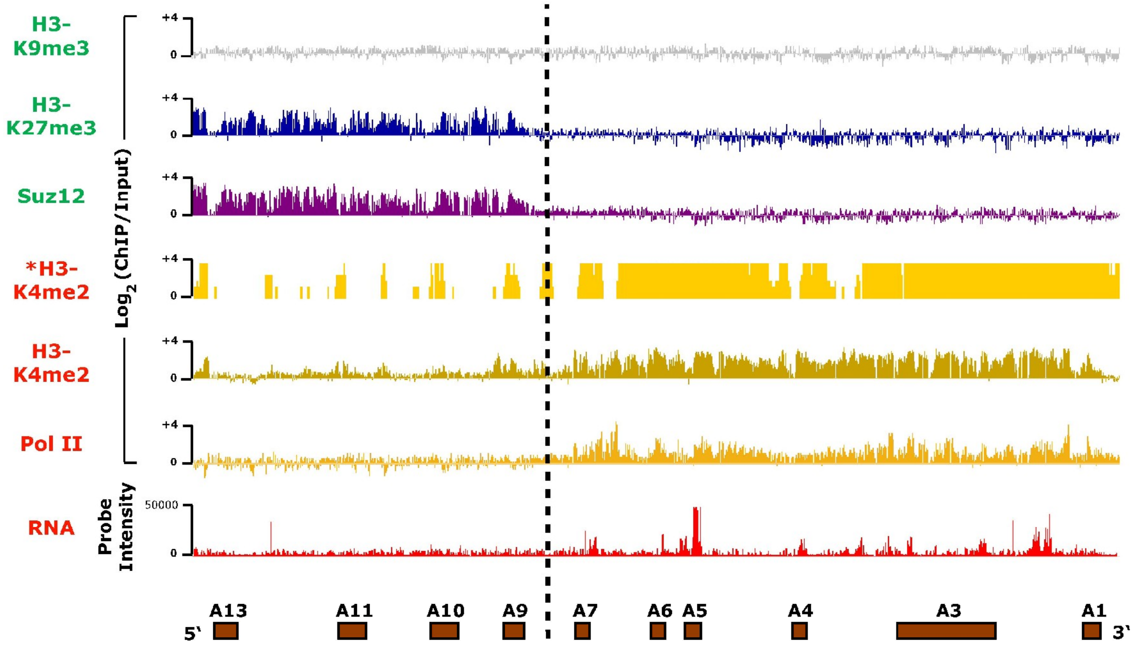

Figure S6. Shown is the ~100 kilobase human HOXA locus. Occupany of PRC2 subunit Suz12, histone H3 trimethylated on lysine 27 (H3K27me3), histone H3 trimethylated on lysine 9 (H3K9-me3), histone H3 dimethylated on lysine 4 (H3K4-me2* = data from Bernstein et al. 2006) and RNA polymerase II (pol II) were determined by chromatin immunoprecipitation followed by tiling array analysis (ChIP-chip). Each ChIP was labeled with Cy5 dye and hybridized against the input chromatin labeled with Cy3 dye. The log2 ratio of each ChIP/Input is plotted on the Y-axis for primary lung fibroblasts.

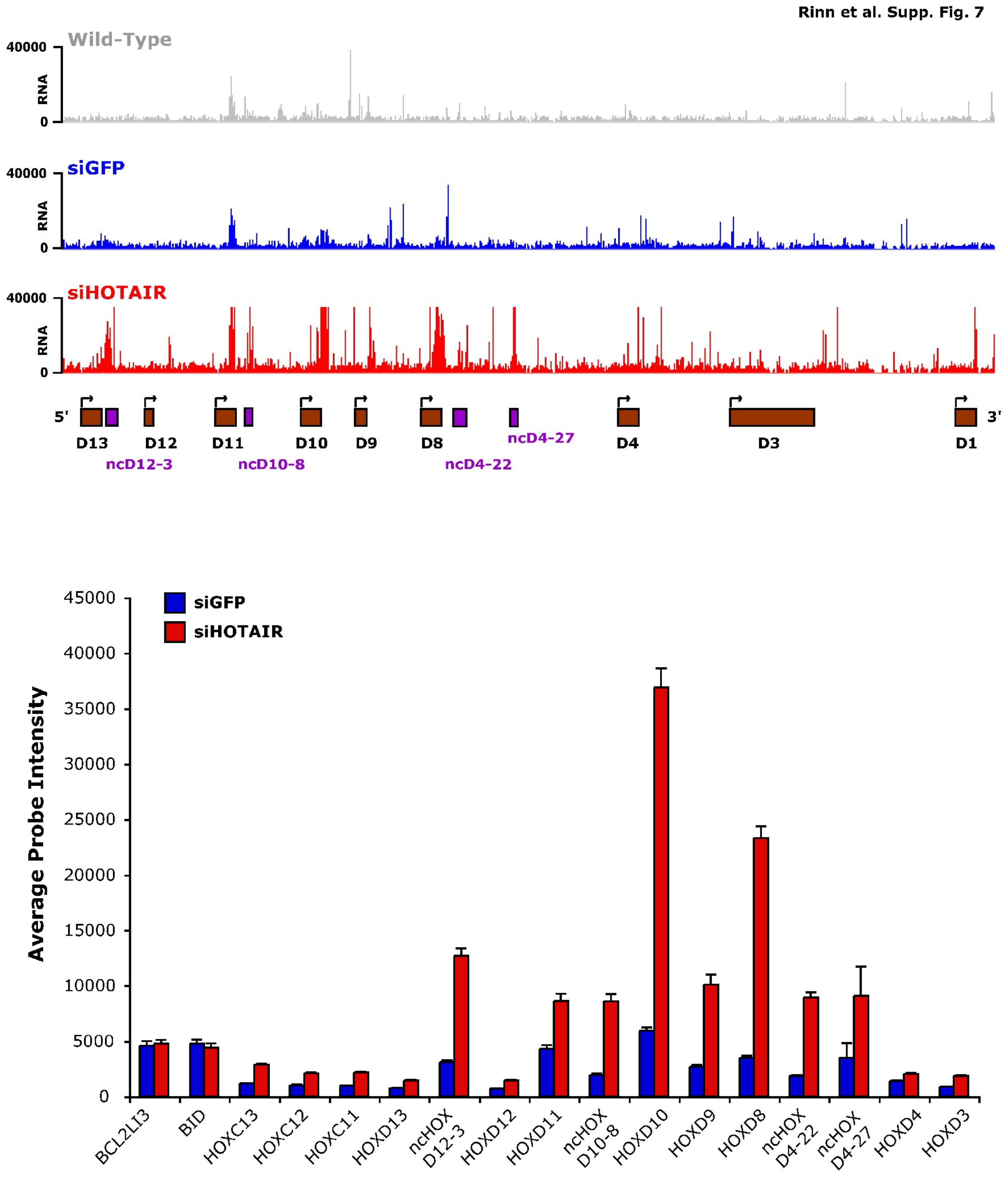

Figure S7.

Top: RNA expression profiles of the HOXD locus for wild-type foreskin cells, foreskin cells treated with RNA interference to GFP (siGFP) and foreskin cells treated with RNA interference to HOTAIR (siHOTAIR). The hybridization intensity of each probe is plotted on the Y-axis on a linear scale. HOXD genes are represented by brown boxes; ncRNAs are shown by purple boxes.

Bottom: Quantitation of RNA levels in HOXD, HOXC, and control chromosome 22 genes after depletion of HOTAIR. The mean ± standard error of hybridization intensities for all probes corresponding to transcribed regions of the indicated genes are shown.

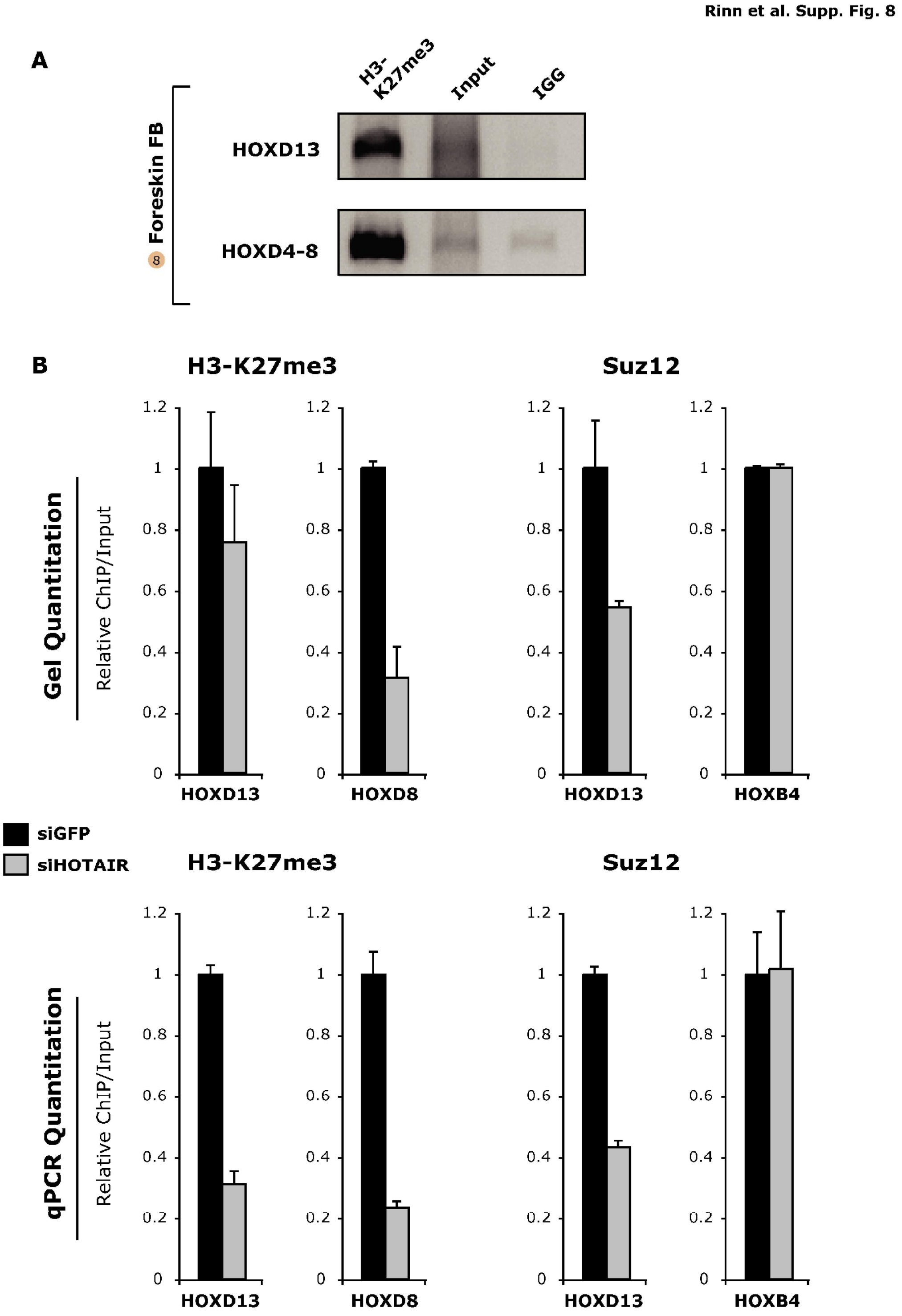

Figure S8. Specificity and quantitation of chromatin immunoprecipitation

(A) Lack of signal in mock ChIP with IgG and enrichment over input for ChIP of H3K27me3 occupancy of HOXD13 promoter and HOXD4-D8 intergenic region in foreskin fibroblasts.

(B) Comparison of ChIP quantitation. Gel quantitation of semi-quantitative PCR analysis (top) and qPCR with SYBR green (bottom) of eight sets of ChIP experiments show similar trends. One exception is the greater fold reduction of H3K27me3 occupancy of *HOXD13* promoter after HOTAIR depletion seen on qPCR.

Rinn et al. Supp. Fig. 9

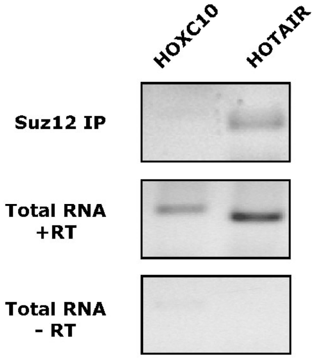

Figure S9. Suz12 does not bind *HOXC10* mRNA.

Native immunoprecipitation of Suz12 in foreskin fibroblasts retrieves HOTAIR ncRNA but not *HOXC10* mRNA. Input RNAs are shown below. HOXC10 was chosen because its genomic location was near HOTAIR and both had similar levels of expression.

SUPPLEMENTAL REFERENCES

Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., and DeRisi, J. L. (2003). Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol *4*, R9.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876-880.

Rinn, J. L., Bondre, C., Gladstone, H. B., Brown, P. O., and Chang, H. Y. (2006). Anatomic demarcation by positional variation in fibroblast gene expression programs. PLoS Genet *2*, e119.

Segal, E., Yelensky, R., and Koller, D. (2003). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics *19 Suppl 1*, i273-282.

Squazzo, S. L., O'Geen, H., Komashko, V. M., Krig, S. R., Jin, V. X., Jang, S. W., Margueron, R., Reinberg, D., Green, R., and Farnham, P. J. (2006). Suz12 binds to silenced regions of the genome in a cell-type-specific manner. Genome Res *16*, 890-900.